Chapter 7

SARS–A post-genomic epidemic

Phylogenetic analysis

- Phylogenetic trees
- The neighbor-joining algorithm
- The Newick format for representing trees

7.1 Outbreak

On February 28, 2003 the Vietnam French Hospital of Hanoi, a private hospital with only 60 beds, called the World Health Organization (WHO) with a report of patients who had unusual influenza-like symptoms. Hospital officials had seen an avian influenza virus pass through the region a few years earlier and suspected a similar virus. The pathogen seemed highly contagious and highly virulent, so they asked that someone from the WHO be sent to investigate. Dr. Carlo Urbani, an Italian specialist in infectious diseases, responded.

Dr. Urbani quickly determined that the Vietnamese hospital was facing a new and unusual pathogen. The infections he observed were characterized by a fever, dry cough, shortness of breath, and progressively worsening respiratory problems. Death from respiratory failure occurred in a significant fraction of the infected patients. For the next several days, Dr. Urbani worked at the hospital documenting findings, collecting samples, and organizing patient quarantine. He was the first person to identify and describe the new disease, called Severe Acute Respiratory Syndrome, or SARS. In a matter of weeks, Dr. Urbani and five other healthcare professionals from the hospital would be dead.

By March 15 the WHO had already issued a global alert, calling SARS a "worldwide health threat." They warned that possible cases had been identified in Canada, Indonesia, Philippines, Singapore, Thailand, and Vietnam.

The origin of the epidemic. Although the origins and cause of SARS were still unknown in March of 2003, it would not be long before the analysis of multiple SARS genomes revealed the story of how this disease had originated and traveled to many countries.

We now know that the first cases of what was to become known as SARS appeared as early as November 2002 in the Chinese province of Guangdong. A few months later the first major outbreak of SARS hit Guangdong: more than 130 infected patients were treated, 106 of whom had acquired the disease while in a hospital in the city of Guangzhou (the rest of the world was unaware of this). A doctor who worked in this hospital visited Hong Kong and checked into the city's Hotel Metropole on February 21, 2003. He was eventually hospitalized with pneumonia during his visit to Hong Kong and died. A number of other travelers staying on the ninth floor of the Metropole became infected and left Hong Kong as disease carriers. One of these was an American businessman named Johnny Chen; he would be the first patient treated in the Vietnam French Hospital of Hanoi (before dying Mr. Chen infected at least 80 people, including half of the hospital workers who cared for him). Other infected travelers from the Metropole would bring SARS to Canada, Singapore, and the United States. By late April 2003, over 4300 SARS cases and 250 SARS-related deaths had been reported to the WHO from over 25 countries around the world. Most of these cases occurred after exposure to SARS patients in household or hospital settings. Many of these cases would later be traced, through multiple chains of transmission, to the doctor from Guangdong province who visited Hong Kong. (On April 5, 2003 China apologized for its slow response to the outbreak.)

In response to the outbreak – and while struggling to control the spread of the pathogen – the WHO coordinated an international collaboration that included clinical, epidemiological, and laboratory investigations to identify the exact cause of the disease, as well as its origin. In the third week of March 2003 laboratories in the United States, Canada, Germany, and Hong Kong independently isolated a novel coronavirus (SARS-CoV) from SARS patients. Evidence of SARS-CoV infection has since been documented in SARS patients throughout the world, indicating that the new virus is in fact responsible for the syndrome.

Coronaviruses are RNA viruses common in humans and animals; they are called coronaviruses because their distinctive halo of spiky envelope proteins resembles a crown. Some of these viruses cause common colds and are responsible for 15–25% of all upper respiratory tract infections, as well as being the cause of important diseases of livestock. In April 2003 scientists from Canada announced the completion of the genome sequence of the SARS virus. *Phylogenetic* analyses revealed the most closely related coronavirus to be one that infected a small mammal (not a bird as initially suspected), the palm civet. Not coincidentally, the palm civet is a part of the diet in the Guangdong province of China.

Phylogenetic analysis of the SARS epidemic. While SARS was still spreading, and television was filled with images of people wearing surgical masks in the streets, scientists were already discussing the significance of the SARS genome. Multiple laboratories around the world were racing to find the origin and ultimately a cure for SARS.

In May of 2003, two papers were published in the journal *Science* that reported the first full genome sequences of SARS-CoV. The agent of the epidemic turned out to be a 29 751 base pair coronavirus that was substantially different from any known human virus; the conclusion, therefore, was that SARS was derived from some non-human virus. Jumping the species barrier is not uncommon for viruses, as quite a number of examples of such zoonotic infections are known (HIV is one other such example). By 2003 the virus had spread to

Africa, India, and Europe, so genome sequences of additional SARS viruses were available from individuals around the world. All of these data were available online, and became a leading example of how virology and medicine can benefit from genomic tools.

Many important questions can be answered by analyzing large sets of complete viral genomes. In this chapter we will present the tools to answer some of them. What kind of virus caused this epidemic? What organism was the original viral host? What was the time and place of the crossing of the species barrier? What are the key mutations that made this switch possible? What was the route of transmission of SARS from the time it crossed the species barrier to its spread around the world?

In order to answer these questions, we will first examine some key algorithms of *phylogenetics*, and then will apply these algorithms to the very SARS data that were obtained in the spring of 2003. (All of these sequences are available from GenBank, and on this book's website.) Phylogenetics – the study of relationships among individuals and species – forms a crucial set of tools in computational genomics, needed for everything from building the tree of life, to discovering the origins of epidemics, to uncovering the very processes that shape genomes. We now take a general look at the algorithms and statistical models that are used to analyze phylogenetic relationships.

7.2 On trees and evolution

The trajectory followed by the SARS virus during the winter and spring of 2003 can be likened to a tree. All of the SARS viruses in the world are related to one another: starting from the single virus that appeared in China, the network of relationships branched over and over again as SARS was passed from person to person. The result of this branching process can be envisioned as a graph with a tree-like structure, and we can infer this tree from the DNA differences between the different SARS genomes (see for example Figure 7.2).

Traditionally, the evolutionary history connecting any group of species or individuals (not just viruses) has been represented by means of a tree, which mathematically is a special type of graph. (We can also represent the relationships among related genes, or other groups using trees; we will call any such units under comparison *taxa*.) We are able to draw such trees because all of the species on earth share a common ancestor, as do all of the individuals within a species. The use of trees to represent relationships among species dates at least to the time of Charles Darwin. Note however that the relation among some genes or species can be more complex than a tree, and we may need to resort to phylogenetic *networks* on these occasions, due to the role of recombination or inter-specific hybridization. In this book we will not explore this important issue further, but will only use phylogenetic trees.

The structure of phylogenetic trees. Whatever taxa we wish to represent the relationships among, a phylogenetic tree attempts to represent both the ordering of relationships (A is closer to B than it is to C), and the distance separating any two groups (i.e. the time since they diverged). The simplest tree contains only two taxa (whether the taxa represent species, individuals, genes, etc.), and is represented in Figure 7.1. The two taxa are called the *external nodes* (or *leaves*), and they are connected by *branches*. Their common ancestor is an *internal node*. In the case of two taxa there is only one possible topology, or ordering of relationships. However, the length of time since the common ancestor can cover a huge range, from one day to one billion years. If both of these taxa are extant (that is, present now), then the time back to the ancestor must be the same for both lineages.

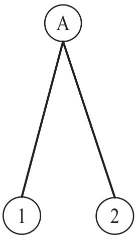


Fig. 7.1 A simple tree with only two leaves: nodes 1 and 2. Node A is the root of the tree

When we have more than two taxa, we must define a number of additional characteristics of trees. Bifurcating trees are those in which every internal node has exactly degree 3 (except for the root node – defined below), and every external node has exactly degree 1. (The degree of a node is the number of branches connecting to it.) Multifurcating trees can have interior nodes with a degree higher than 3. In evolutionary trees, the external nodes represent existing taxa, while the internal nodes represent their ancestors (typically, but not necessarily, extinct). Bifurcating trees can thus be thought of as requiring that every ancestor leads to only two descendants. Note that we normally assume that the true tree is bifurcating, that one ancestor leads to only two descendants at splitting; however, we can sometimes use multifurcating trees to represent uncertainty in the order of splitting events.

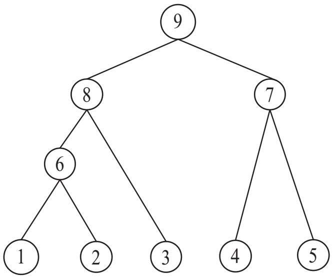


Fig. 7.2 A bifurcating rooted tree with five leaves, three internal nodes, and one root

Rooted vs unrooted trees. We can also define a phylogenetic tree as either *rooted* or *unrooted*. In a rooted tree we define a special internal node called the root. The root node is the common ancestor to all the other nodes in the tree and all evolutionary paths lead back to the root (the root node has a degree of 2, as can be seen in Figure 7.2). When we have a rooted tree, we can consider its branches to have an orientation going from the root to each external node. Unrooted trees, on the other hand, are un-oriented; they show the topological relationships among taxa, but are agnostic with respect to the identity of the common ancestor of all taxa (see Figure 7.3).

The task of finding the root (choosing an edge of an unrooted tree where to place the root node) requires external biological information, or at least some assumptions about where the root should be placed. The root is typically defined by including in the dataset one or more taxa that are known to be the result of an earlier split, and hence to be more distantly related to each of the other taxa. This external taxon (or taxa) is called an *outgroup*. The branch of the tree

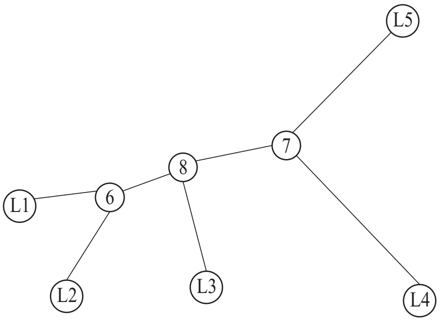


Fig. 7.3 An unrooted tree with five leaves and three internal nodes

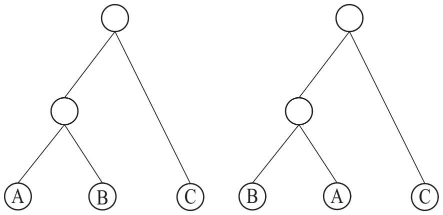


Fig. 7.4 These two trees are mathematically different, but they are biologically equivalent: the order in which the children of a node are represented is irrelevant, and hence phylogenetic trees are to be considered invariant with respect to changes of this order

where the outgroup joins the rest of the taxa then is considered to contain the root node. Additional methods for defining the root based on the structure of the tree are also used, and we discuss some of them below.

Finally, we must underscore an important component of these graphs: any rotation of branches about an internal node does not change the relationships among taxa. No matter the exact order in which we list the taxa from left to right, the tree represents the same set of phylogenetic relationships (see for example Figure 7.4).

The number of possible trees. The algorithmic and statistical problem of reconstructing a phylogenetic tree from a set of related DNA sequences is greatly complicated by the huge number of possible trees. The number of possible tree topologies is a function of n, the number of taxa in the tree, and depends on whether a tree is rooted or unrooted. The number of unrooted trees for $n \geq 3$ is

$$\frac{(2n-5)!}{2n-3(n-3)!}.$$

The number of rooted trees for $n \geq 2$ is

$$\frac{(2n-3)!}{2n-2(n-2)!}.$$

We can see that there is only one possible unrooted tree for the case of three taxa; the same holds for the case of two taxa in a rooted tree. For either kind of tree, the number of possibilities increases very quickly as n rises. For five taxa there are already 105 rooted trees, and for ten taxa there are 34 459 425 rooted trees. Only one of these is the true tree of evolutionary relationships, and inferring it from data is one of the main tasks of phylogenetic analysis.

Representing trees. There are various ways to non-graphically represent a tree, $\mathcal{T}$. A simple representation is – as with other graphs – by listing its nodes and the neighbors of each node. For oriented trees (those with a root) it is even easier: we only need to list the descendants of each node. So if we number the external nodes from 1 to n, for a bifurcating tree, there will always be $n-1$ internal nodes, and hence a matrix of all internal nodes and their children will be $(n-1) \times 3$ and will suffice to fully specify the tree topology (see the next example). Note that in this chapter we will denote by $\mathcal{T}$ both the graph itself and its representation in one of the various equivalent non-graphical formats.

Example 7.1
Tree representation as an array. The topology of the 5-taxa rooted tree in Figure 7.2 can be represented by the array:

9	8	7
8	3	6
7	4	5
6	1	2
5	–	–
4	–	–
3	–	–
2	–	–
1	–	–,

where each entry in the left-most column represents a node (internal and external), and the two corresponding entries on the same row represent its children. External nodes (here 1–5) have no descendant nodes, and hence the last five rows could be omitted. Note that this representation only works for rooted trees and is not the most intuitive one for biological interpretation; however, it is often used in implementation. If we also want to specify the distance between nodes (the branch lengths), we can either add this information to the above matrix, in the two right-most columns, or create a separate array that contains the distance from each node to its direct ancestor.

The example above shows one of the many possible representations of a tree. Another, more intuitive, representation of directed trees exploits the relationship between parental nodes and their descendants, using parentheses. In this representation, the tree $\mathcal{T}$ would be represented as

$$(((1, 2), 3), (4, 5)).$$

This is the same tree that appears in Figure 7.2. A popular standard tree file format called *Newick Format* is based on this idea, and is described in Section 7.5.

7.3 Inferring trees

7.3.1 Introduction to phylogenetic inference

As we have seen, the relationships among organisms can be viewed as a tree, and this representation is likely to be a very close, if not exact, model for the true relationships among species, individuals, and even genes. But the underlying, true tree is unknown. Yes, in the case of SARS we might have been able to re-create the true tree if we had known of the original infection and tracked its passing from one person to the next; but more often than not we are simply presented with organisms in the present day and asked to infer the most likely set of relationships connecting them. Just a few years ago this task would have been done largely by tracking changes in morphological characters: visible differences in the organisms that might tell us about their underlying genetic relatedness. Starting in the 1980s, and concomitant with technological advances in DNA sequencing technologies, inferring phylogenetic trees became a task inextricably linked to the analysis of DNA. More recently, it has been based on whole-genome comparisons.

There is always a true tree (or tree-like diagram) that describes the relationships among organisms. This unknown tree can be inferred from a comparison of the DNA sequences of these organisms because the sequences are always changing, leaving behind a trail of mutations that will be present in the descendants of mutant genes and absent from all other individuals. (Note that this line of reasoning relies heavily on the fact that mutation is rare, and will therefore lead unrelated individuals to the same sequence extremely infrequently.) If a gene or any segment of DNA did not change over time, we would have no record of its past. But mutation ensures that there is a traceable history of relationships.

Within a set of organisms we expect that every gene that they share will lead us to the same or very similar trees. Each of the genes might mutate and evolve at a different rate, but all of the genes will be inherited as a group and will be passed to descendants together, resulting in the same tree. Recombination between sequences within a species can cause two genes (or different parts of the same gene) to have different histories, but as we said earlier we will ignore this complication for now. So regardless of the exact DNA sequence we choose to examine, we expect to obtain the same tree from our analyses. And the more sequence we examine – and hence the more mutations – the more power we gain to resolve relationships between closely related organisms. However, although the true tree reflects the fact that the organisms at the external nodes are all equally distant from their common ancestor (in terms of time), different genes evolve at different rates, and different species may even have different mutation rates. As a result, though all of the external taxa in the true tree are the same distance from the root node, the vagaries of mutation mean that inferred trees may include some very long branches and some very short branches. All of the external taxa may therefore not necessarily be the same measured genetic distance from the root (a fact that can be very interesting for the study of changes in mutation or substitution rates).

Given a set of taxa and homologous sequences from each of them, there are a number of common ways to reconstruct their phylogenetic relationships. The methods can be broadly divided into two groups: those that rank all possible trees using some criterion in order to find the optimal one; and those that directly build the tree from the data (without explicitly stating a scoring function). Within the first group, criteria used often center around finding the tree with the smallest number of necessary mutations to explain the data (via likelihood and other methods). Given the huge number of possible trees, these methods can take a very long time to find the best tree, and even then may not necessarily find this tree because of the approximations that must be used to speed up the search. However, likelihood-based methods are favored for in-depth phylogenetic analyses. The second group includes phylogenetic methods that are themselves both criteria and algorithms for building trees, and are often based on computing the pairwise distance between taxa as a first step. They are typically very fast and have hence become extremely popular in genomic analyses. Although not necessarily as well behaved statistically as other methods, the most popular of these so-called "distance" methods (the *neighbor-joining* algorithm) is surprisingly robust and accurate. It is guaranteed to infer the true tree if distances used reflect the true distances between sequences, a result that is often not guaranteed by more statistically sophisticated methods.

7.3.2 Inferring trees from distance data

For a set of n taxa $\{\tau_1, \ldots, \tau_n\}$, we represent the genetic distance between them as a matrix of pairwise distances, D, usually taken from pairwise alignments and corrected for multiple substitutions by schemes such as Jukes–Cantor (with each distance measure between taxa given by $d(i, j)$). Given a distance matrix, simple and efficient algorithms can be used to infer the tree as long as these distances are *additive* (a notion defined shortly below) and often even when they are not.

Additivity and distance matrices. If the branches within a tree each have a specified length, then the distance between any two nodes can easily be computed as the total length of the (unique) path connecting them (see Figure 7.5 for an example).

In this way a tree can specify a distance matrix between its leaf nodes. However, not all distance matrices have this "additivity" property. Intuitively, the notion of additivity means that a certain distance can be represented by a tree. Formally, this translates into a technical condition that is stated below, and that motivates the algorithmics used in this chapter. Biologically, additivity is an important property for a distance matrix: the actual number of substitution events separating two taxa from their last common ancestor (their genetic distance) forms an additive distance. One of the attractions of using substitution models such as Jukes–Cantor is that they attempt to make the distance matrix more additive, and hence they make inference of the tree easier.

We can represent the distances within the tree of Figure 7.5 with the following distance matrix:

	L1	L2	L3	L4	L5
L1	0	2	4	6	6
L2	2	0	4	6	6
L3	4	4	0	6	6
L4	6	6	6	0	4
L5	6	6	6	4	0

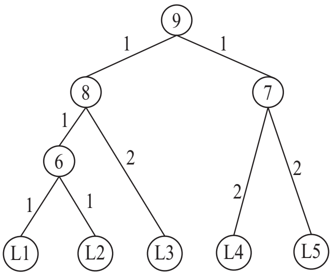


Fig. 7.5 A rooted tree with length-annotated branches.

Definition 7.1
Additive tree of a distance matrix. Let D be a symmetric $m \times m$ matrix where the numbers on the diagonal are all zero and the off-diagonal numbers are all strictly positive. Let $\mathcal{T}$ be an edge weighted tree with at least m nodes, where m distinct nodes of $\mathcal{T}$ are labeled with the rows of D. Tree $\mathcal{T}$ is called an *additive tree* for matrix D if, for every pair of labeled nodes (i, j), the path from node i to node j has total weight (or distance) exactly $d(i, j)$.

Definition 7.2
Length of a tree. We also define the *total length* of a tree as the sum of all the branch lengths.

7.3.3 The neighbor-joining algorithm

The most popular distance-based method for inferring phylogenetic trees is known as *neighbor-joining*. The neighbor-joining (NJ) algorithm was first described by Naruya Saitou and Masatoshi Nei in 1987 (this is the same Nei as in the Nei–Gojobori method discussed in Chapter 6). NJ is a greedy algorithm that starts with an initial star phylogeny (one in which all taxa are connected directly to a single root node) and proceeds by iteratively merging pairs of nodes. The criterion with which each pair of nodes is selected is the key to its success: it identifies nodes that are topological neighbors in the underlying tree by using a mathematical characterization that is valid for all additive distance matrices.

After selecting the two taxa, they are merged into a single taxon, which will be further treated as a new single taxon. A new modified distance matrix is then created, where the distances of other taxa to the composite taxon are calculated. This process is repeated until all of the taxa have been joined together. Because NJ produces unrooted trees, an outgroup or other method is needed to specify the root node. As noted earlier, if the distance matrix used to build the tree is additive, then NJ will give the true tree; if, however, it is non-additive (i.e. there is noise in the data), there can be ambiguities in the inferred tree. Below we provide details for the calculation and construction of trees with NJ (using the corrected method published by James Studier and Karl Keppler), including some of the necessary background that makes it clear how NJ works. The reader can safely skip this technical part if they prefer.

Finding branch lengths. The lengths of individual branches in an unrooted tree with three external taxa can be computed from pairwise distances for additive matrices. To see this, imagine first that we have three taxa in an unrooted tree with known distances on all branches. Consider this unrooted tree joining taxa A, B, and C, with the length of each branch L_x, L_y, and L_z respectively (see Figure 7.6); then:

$$L_x + L_y = d_{AB}$$
$$L_x + L_z = d_{AC}$$
$$L_y + L_z = d_{BC}.$$

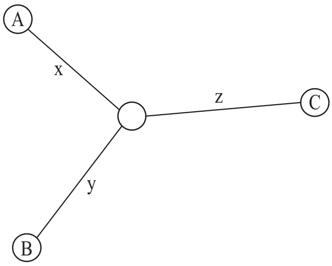


Fig. 7.6 Illustration of the 3-point formula

The solution for this system is

$$L_x = \frac{(d_{AB} + d_{AC} - d_{BC})}{2}$$
$$L_y = \frac{(d_{AB} + d_{BC} - d_{AC})}{2}$$
$$L_z = \frac{(d_{AC} + d_{BC} - d_{AB})}{2}.$$

This formula is called the *3-point formula*, and shows how we can infer the individual branch lengths on a tree from a set of pairwise distances.

A test of neighborliness for two taxa. Neighbor-joining proceeds by finding neighbors in the tree, and iteratively joining them until the whole tree is complete. In order to do this, NJ must be able to identify these neighboring taxa. It is a remarkable property of additive distances that it is possible to devise a test to find nodes that are neighbors in the underlying tree.

Consider two taxa, τ_1 and τ_2, that are joined to the vertex V and two other generic neighbor nodes τ_i and τ_j (as in Figure 7.7). Then the following inequality holds when τ_1 and τ_2 are neighbors:

$$d(\tau_1, \tau_2) + d(\tau_i, \tau_j) < d(\tau_1, \tau_i) + d(\tau_2, \tau_j).$$

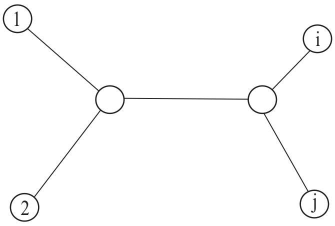


Fig. 7.7 The 4-point condition, used in the development of the neighbor-joining algorithm

This can be seen from Figure 7.7. In words, the sum of the distances between τ_1, τ_2, τ_i, and τ_j should be minimized when neighbors are paired in the summation. This leads to a criterion for detecting neighbors when there are an arbitrary number of external nodes in a tree. First, define the total distance from taxon τ_i to all other taxa as

$$R_i = \sum_{j=1}^{n} d(\tau_i, \tau_j),$$

where the distance $d(\tau_i, \tau_i)$ is naturally interpreted as 0. We also define the "neighborliness" between two taxa to be

$$M(\tau_i, \tau_j) = (n-2)d(\tau_i, \tau_j) - R_i - R_j, \tag{7.1}$$

where we are minimizing both the distance between external nodes and the total distance in the tree. For any two nodes τ_i and τ_j that are neighbors, we require that

$$M(\tau_i, \tau_j) < M(\tau_i, \tau_k)\ \forall k \neq j.$$

This gives us the crucial piece of the NJ algorithm: from the distance matrix D that contains all of the pairwise distances $d(\tau_i, \tau_j)$, compute a new table of values for $M(\tau_i, \tau_j)$. Then choose to join the pair of taxa with the smallest value of $M(\tau_i, \tau_j)$. We call this the *4-point condition.*

Joining neighbors in the tree. The NJ algorithm chooses to merge two nodes that satisfy the criterion above into a new node, V, and then computes the distance between V and all of the other nodes, as well as the length of the newly created branches to τ_i and τ_j using the 3-point formula. It then recomputes M (now using V as an external taxon) and iterates.

We calculate the distance from the new node V to each of the remaining external nodes as

$$d(V, \tau_k) = \frac{1}{2}[d(\tau_i, \tau_k) + d(\tau_j, \tau_k) - d(\tau_i, \tau_j)]\ \mathit{for}\ k \neq i, j. \tag{7.2}$$

The 3-point formula described above gives us the branch lengths from V to the joined neighbors τ_i and τ_j. The branch lengths from τ_i to V and τ_j to V are given by

$$L(\tau_i, V) = \frac{d(\tau_i, \tau_j)}{2} + \frac{R_i - R_j}{2n-2} \tag{7.3}$$

$$L(\tau_j, V) = \frac{d(\tau_i, \tau_j)}{2} + \frac{R_j - R_i}{2n-2}. \tag{7.4}$$

Below we summarize the various steps of the algorithm.

Rooting the tree. The procedure described above constructs an unrooted tree. In order to define a root node, we need to specify an outgroup, so that the root is assumed to be on the branch connecting the outgroup to the rest of the tree. The midpoint of this branch is a possible choice for the root, but other criteria may be used, for example some aimed at producing a more balanced tree. Here we simply apply so-called midpoint rooting to all of the trees we construct.

NJ algorithm. Input: an $n \times n$ distance matrix, D, and the specification of an outgroup.
Output: a rooted phylogenetic tree, $\mathcal{T}$, represented by a table of relationships, T, and a separate array with all branch lengths, TD:

Step 1: Given a pairwise distance matrix for n taxa, D, compute a new table of values of $M(\tau_i, \tau_j)$ as defined in equation (7.1). Choose the smallest value in this matrix to determine which two taxa to join.

Step 2: If τ_i and τ_j are to be joined at a new vertex V, first calculate the distance from V to the remaining external nodes using equation (7.2). Use these values to update the distance matrix, D, replacing τ_i and τ_j by V.

Step 3: Compute the branch length from τ_i to V and τ_j to V using equations (7.3) and (7.4). Set the values of $T(V, 1) = \tau_i$, $T(V, 2) = \tau_j$, $TD(\tau_i) = L(\tau_i, V)$ and $TD(\tau_j) = L(\tau_j, V)$. These describe the tree topology and branch lengths.

Step 4: The distance matrix now includes $n - 1$ taxa. If there more than two taxa remaining, go back to step 1. If only two taxa remain, join them by a branch of length $d(\tau_i, \tau_j)$.

Step 5: Define a root node on the branch connecting the outgroup to the rest of the tree.

UPGMA. The NJ algorithm reduces to a simpler method, called UPGMA (Unweighted Pair Group Method with Arithmetic Averages), in the case when the matrix M is defined to be equal to the distance matrix D. This means that the distance from the leaf taxa to the root is the same for all taxa, a condition called *ultrametricity*. While this condition must hold for the true tree, in practice it is almost never true of DNA sequence data and therefore leads to erroneous inference of phylogenetic trees. However, UPGMA was an early distance-based phylogenetic method, and is related to the hierarchical clustering algorithm discussed in Chapter 9.

7.4 Case study: phylogenetic analysis of the SARS epidemic

7.4.1 The SARS genome

The genome sequence of SARS-CoV obtained by the Canadian group in April 2003 is a 29 751 bp, single stranded RNA sequence. It can be accessed via GenBank (accession number AY274119.3), and a map of its genes is provided in

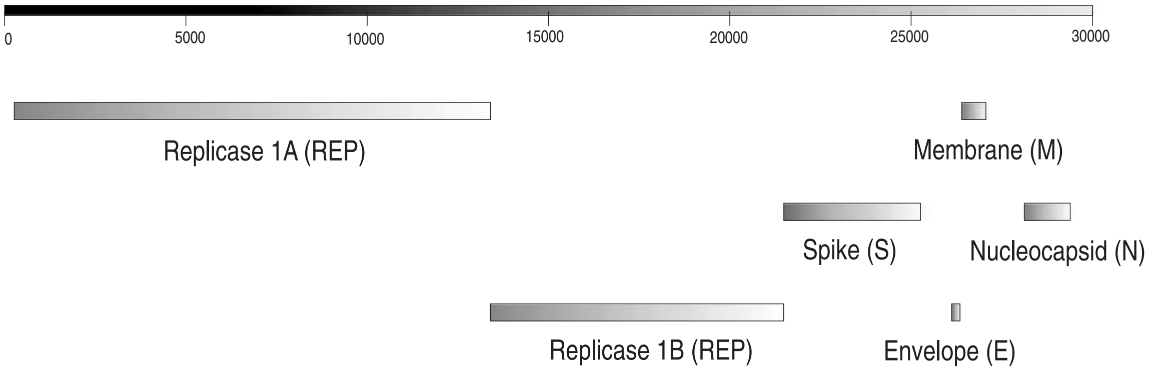


Fig. 7.8 The ORF map of SARS genome. Two-thirds of the genome contains the gene for the replicase protein, the remaining third various key genes, including spike and envelope. The order of these genes is typical of other coronaviruses

Figure 7.8. Its GC content is approximately 41%, within the range for published complete coronavirus genome sequences (37–42%). It has a structure typical of coronaviruses, with five or six genes in a characteristic order. Using the methods described in Chapter 3 we can easily find most of these genes. Notice however that – as with HIV – things can be a little more complicated in viruses, with overlapping ORFs and other problems that can prevent simple ORF-finding methods from identifying all coding regions.

7.4.2 Reconstructing the epidemic

During the SARS epidemic, many of the key questions about its origin and nature could be answered by genomic sequence analysis. The sequence of SARS was obtained and published by various groups in early 2003, and was used as the basis for investigation into the origin and spread of the epidemic. The entire epidemic can now be re-created with the many viral sequences available in GenBank. By building a phylogenetic tree of the isolates from known dates and places, we can observe the crucial role played by the Hotel Metropole. We have chosen 13 sequences for which we could find date and location of the sample (see Table 7.1), and use them to demonstrate how sequence information can illuminate the unfolding of an epidemic.

Identifying the host. The SARS virus was recognized early on as a coronavirus, having the same genes in the same order as other known coronaviruses. It was, however, very different from any other known human coronavirus, and hence its origin was likely to be from another animal. A NJ tree of the spike proteins for various animal coronaviruses, including the coronavirus found in the palm civet, leaves little doubt about the closest relative of SARS coronavirus. SARS appears most closely related to the Himalayan palm civet coronavirus (Figure 7.9), and is quite distantly related to other human coronaviruses.

The disease was not carried by birds but it originated in the palm civet, and later adapted to be spread from human to human.

The epidemic tree. Using the 13 genomes in Table 7.1, a neighbor-joining tree of the spike protein was constructed (see Figure 7.10). The distance matrix was obtained by Jukes–Cantor corrections on genetic distance calculated from global alignments of the spike nucleotide sequence.

Table 7.1 Name, location, and sampling date of SARS virus isolates used in our case study

Name of isolate	Acc. number	Date	Location
GZ01	AY278489	DEC-16-2002	Guangzhou (Guangdong)
ZS-A	AY394997	DEC-26-2002	Zhongshan (Guangdong)
ZS-C	AY395004	JAN-04-2003	Zhongshan (Guangdong)
GZ-B	AY394978	JAN-24-2003	Guangzhou (Guangdong)
HZS-2A	AY394983	JAN-31-2003	Guangzhou Hospital
GZ-50	AY304495	FEB-18-2002	Guangzhou (Guangdong)
CUHK-W1	AY278554	FEB-21-2003	Hong Kong
Urbani	AY278741	FEB-26-2003	Hanoi
Tor 2	AY274119	FEB-27-2003	Toronto
Sin2500	AY283794	MAR-01-2003	Singapore
TW1	AY291451	MAR-08-2003	Taiwan
CUHK-AG01	AY345986	MAR-19-2003	Hong Kong
CUHK-L	AY394999	MAY-15-2003	Hong Kong

We can read the entire story of the epidemic on this tree. Using the palm civet as an outgroup, we see that all the early cases occurred in the Guangdong province, and that the Hotel Metropole coronavirus is almost identical to one of these sequences (i.e. there is no discernible genetic distance between them; Figure 7.10). The rest of the world-wide epidemic is seen to be nested within these initial infections: the events in Singapore, Hanoi, Taiwan, and Toronto can all be traced to the Hong Kong hotel and/or Guangdong. (The sequence of the Hanoi coronavirus is now called the Urbani strain.) The main question that remains is: When did the epidemic start? The answer can be found in the data from Table 7.1.

Date of origin. Because we know the date of collection of each of the SARS viruses for which we have sequence, we are able to observe the progression of mutations over time. For convenience, we again use the ORF corresponding to the spike protein. Relative to the sequence from the palm civet, we see that genetic distance (the y-axis of Figure 7.11) increases with time, in a roughly linear manner (time is along the x-axis, with the 0 point representing January 1, 2003). If we interpolate a least-squares line through these data, we can estimate the approximate date for the origin of the epidemic. Any date compatible with zero distance from the palm civet is a plausible start date for the epidemic, and we estimate it to be an interval centered around September 16, 2002 (106 days before January 1, 2003). The 95% confidence intervals are also shown in Figure 7.11. The method we have used is rather crude, and relies on many assumptions that we cannot verify, yet it delivers a very plausible estimate since the earliest reported cases can be traced back to the second half of 2002.

Area of origin. Even though we now know that the Guangdong province was the area of origin of the epidemic, we can use the same method as presented in Chapter 5 for the origin of humans to look for the likely area of origin of

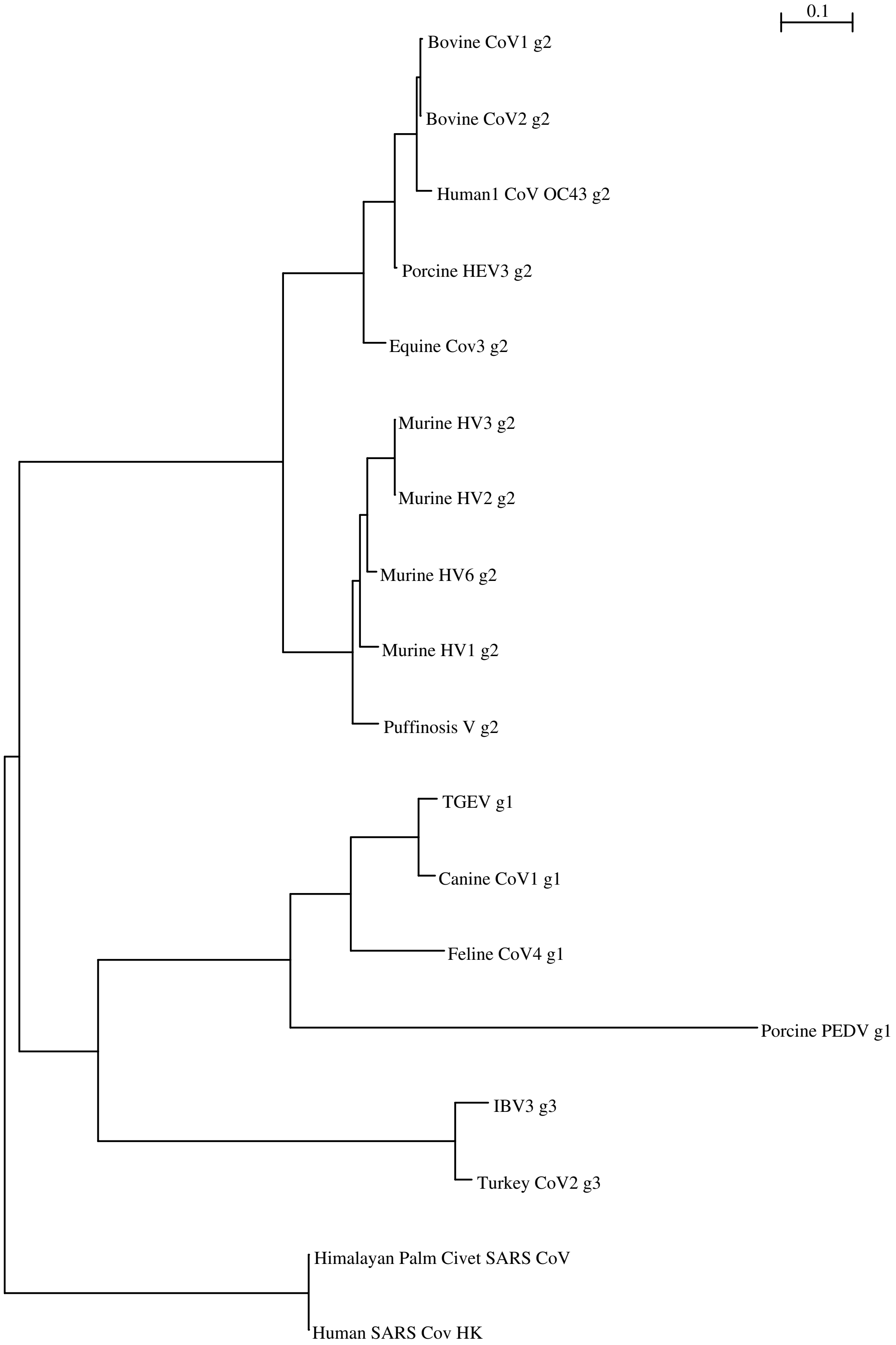
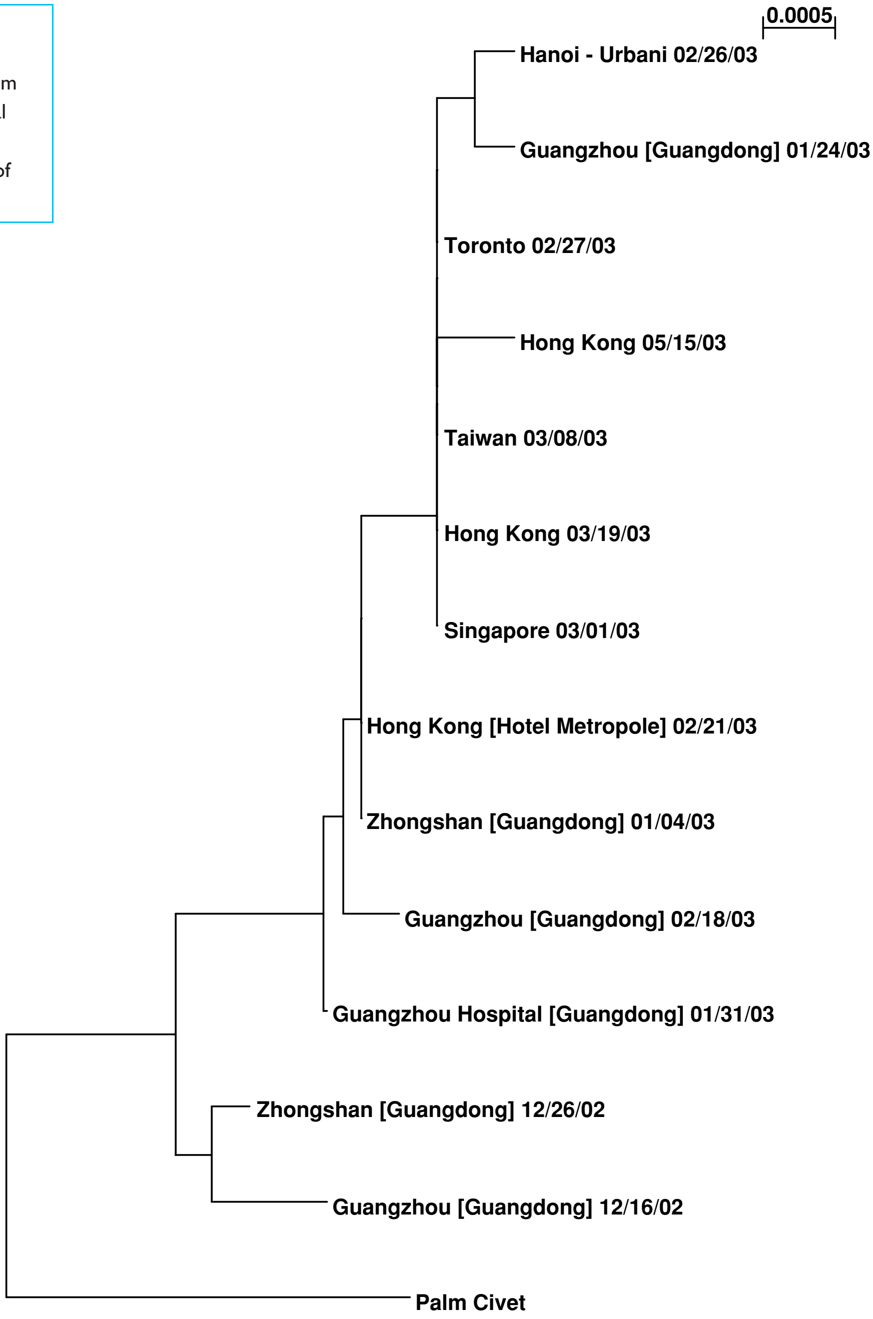


Fig. 7.9 Phylogenetic tree connecting various coronaviruses. Clearly the closest relative to SARS is not a bird virus, but that of palm civet

Fig. 7.10 The story of this epidemic can be read on the phylogenetic tree obtained from genomic data. Note the crucial position of the sequence from Hotel Metropole as the root of the international epidemic


0.0005
Hanoi - Urbani 02/26/03
Guangzhou [Guangdong] 01/24/03
Toronto 02/27/03
Hong Kong 05/15/03
Taiwan 03/08/03
Hong Kong 03/19/03
Singapore 03/01/03
Hong Kong [Hotel Metropole] 02/21/03
Zhongshan [Guangdong] 01/04/03
Guangzhou [Guangdong] 02/18/03
Guangzhou Hospital [Guangdong] 01/31/03
Zhongshan [Guangdong] 12/26/02
Guangzhou [Guangdong] 12/16/02
Palm Civet

SARS. We again take the high nucleotide diversity between sequences in Guangdong as an indication that the virus originated there, with the lower diversity outside this area a result of the subsequent international spread of the single Hotel Metropole strain. Using the genetic distance matrix, we plot (by multidimensional scaling) each sequence as a point. Figure 7.12 shows clearly that there is more diversity among the Guangdong sequences than among all the sequences collected abroad. Of course the tree of Figure 7.10 is another way

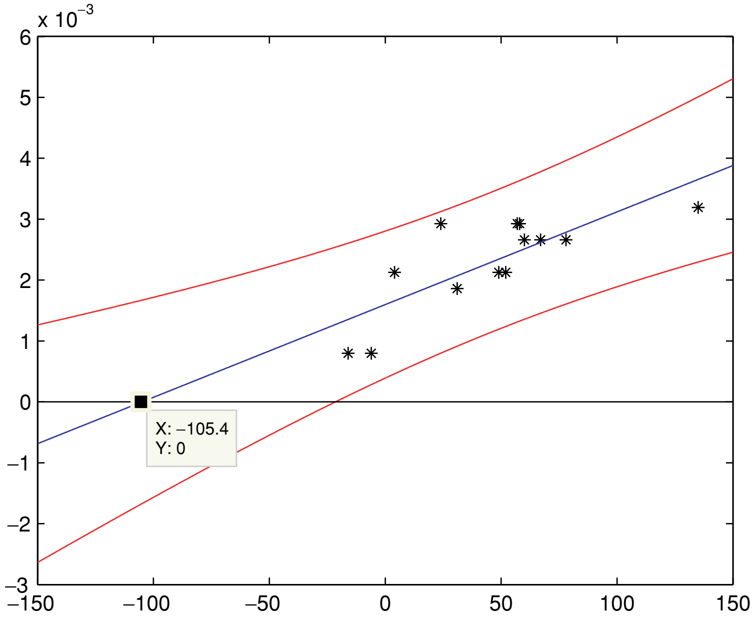
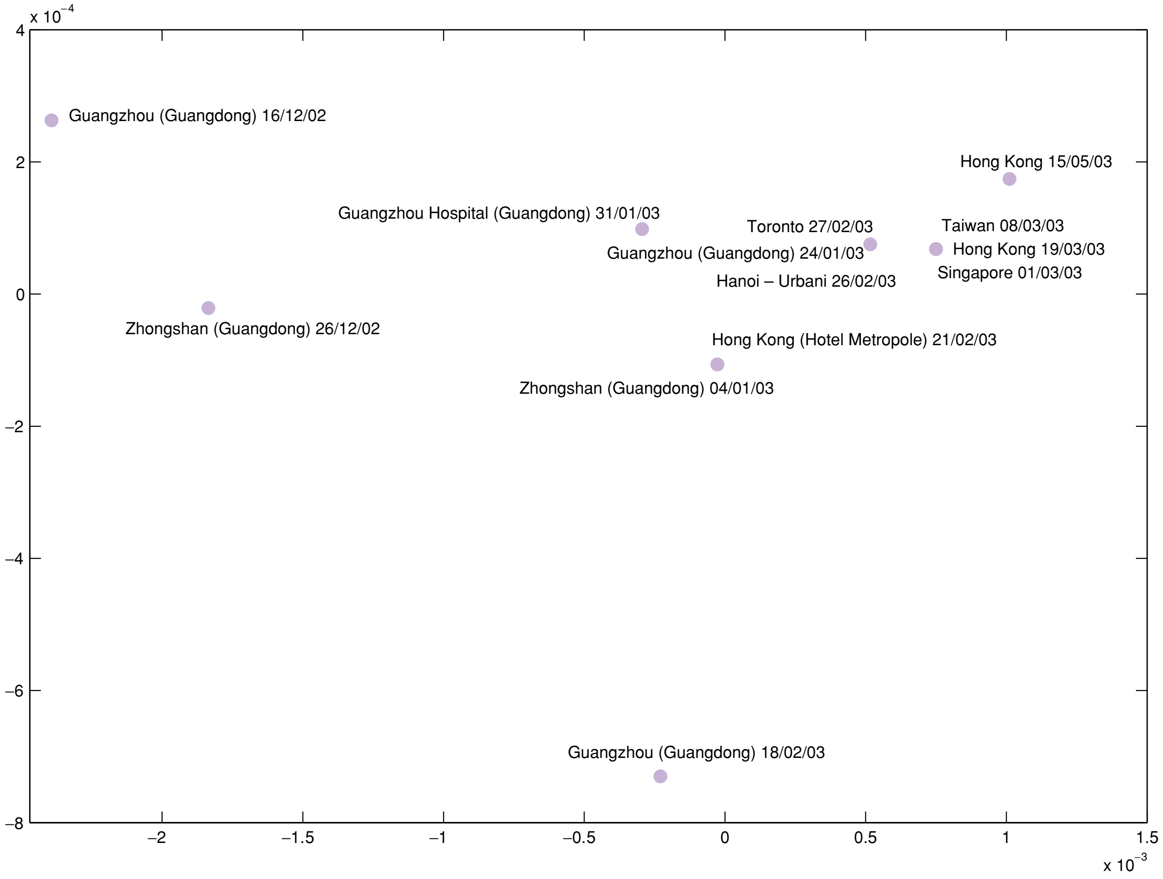


Fig. 7.11 The genetic distance of samples from the palm civet increases roughly linearly with time

Fig. 7.12 Multidimensional scaling visualization of the genetic distance between spike genes shows that the international epidemic sequences are not as diverse as those found in the Guangdong province of China



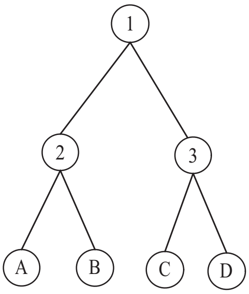


Fig. 7.13 The Newick representation of this tree is ((A, B), (C, D));

of expressing the same information by looking at the lengths of branches separating the various viruses. Both the phylogenetic tree and the multidimensional scaling therefore strongly suggest that Guangdong Province was the area of origin for SARS.

7.5 The Newick format

As an alternative to the rather awkward matrix tree representation introduced in the beginning of this chapter, the Newick format makes use of the correspondence between trees and nested parentheses. This becomes very useful for representing trees in computer-readable form. For example, the tree in Figure 7.13 can be represented in the previously introduced format by the array in (7.5):

$$\begin{array}{ccc} 1 & 2 & 3 \\ 2 & A & B \\ 3 & C & D\,. \end{array} \tag{7.5}$$

In the Newick format this tree is represented by the following sequence of printable characters:

((A, B), (C, D));

the convention is that the tree file must end with a semicolon. Interior nodes are represented by a pair of matched parentheses. Between them are representations of the nodes that are immediately descended from that node, separated by commas. In the above tree, the immediate descendants of the root are two interior nodes. Each of them has two descendants. In our example these happen to be leaves, but in general they could also be interior nodes and the result would be further nesting of parentheses. Leaves are represented by their names. A name can be any string of printable characters, except blanks, colons, semicolons, parentheses, and square brackets. Any name may also be empty; a tree like:

((,), (,));

is allowed in the file format. Note also that trees can be multifurcating; that is, nodes can have more than two children.

Branch lengths can be incorporated into a tree by putting a real number, with or without decimal point, after a node and preceded by a colon. This represents the length of the branch immediately above that node. Thus the above tree might have lengths represented as

((A : 1.0, B : 1.0) : 2, (C : 1, D : 1) : 2);

The tree starts on the first line of the file, and can continue to subsequent lines. It is best to proceed to a new line, if at all, immediately after a comma. Blanks can be inserted at any point except in the middle of a species name or a branch length. Names can also be assigned to interior nodes: these names follow the right parenthesis for that interior node, as in

((A, B)2, (C, D)3);

7.6 Exercises

(1) Using the mtDNA data discussed in Chapter 5, create a tree comparing human and apes. Discuss the position of Neanderthal. Compare the trees obtained with and without the Jukes–Cantor correction.

(2) The mysterious origins of HIV, described in Chapter 5, have been the subject of many investigations. In particular, the relation between HIV and a similar virus found in monkeys, SIV, has been debated and studied in depth. Genomic analysis of various strains of HIV and SIV can be used to settle the question. The data are available on the book's website, both for complete genomes and just for the ENV protein. Using the free package Phylip, construct the phylogenetic tree of various strains of HIV and SIV. This tree should show that the virus crossed the species barrier twice, once leading to the HIV1 epidemic and the second time leading to the HIV2 epidemic. The book's website contains a reconstruction of the tree as a reference.

(3) Construct the tree of the odorant receptors and related proteins, as discussed in Chapter 4 (dataset available on the book's website).

7.7 Reading list

The genome sequence of the SARS-associated coronavirus was first reported in Marra *et al.* (2003) and Rota *et al.* (2003) and its evolution discussed in Guan *et al.* (2003) and Consortium (2003). The timing of the last common ancestor of SARS viruses can be found in Zeng *et al.* (2003) and Lu *et al.* (2004) and the estimation of the area of origin is discussed in Zhang and Zheng (2003). Its phylogeny with other coronaviruses is presented in Eickmann *et al.* (2003) and in Lei *et al.* (2003), and a more general discussion of viral evolution can be found in Holmes and Rambaut (2004). The story of Carlo Urbani's death can be found in Reilley *et al.* (2003).

A complete textbook on phylogenetic tree inference is the excellent Felsenstein (2004), which should be considered the starting point of any investigation into the algorithmic and statistical issues concerning computational phylogenetic analysis. A classical reference on the construction of phylogenetic trees is Fitch and Margoliash (1967). The neighbor-joining algorithm was introduced in Saitou and Nei (1987).

Interesting introductory readings on phylogenetic topics are Doolittle (2000) (on the tree of life) and Cann and Wilson (2003) (on human origins). A discussion of the phylogenetic tree of HIV can be found on the book's website.

Freely available packages for phylogenetic analysis include CLUSTALW Thompson *et al.* (1994) and Phylip, the phylogeny inference package created by J. Felsenstein (Felsenstein, 2004). Very popular tree visualization tools include NJPLOT, UNROOTED, and TreeView, whose complete references and web coordinates are available via the book's website.

Links to software packages, and to all of the above-mentioned papers, datasets, and websites, can be found on the book's website:

`www.computational-genomics.net`